日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　１月３１日

出願番号
Application Number: 特願２００３－０２３３４２

[ST. 10/C]: [ＪＰ２００３－０２３３４２]

出願人
Applicant(s): 松下電器産業株式会社

２００４年　３月　４日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫

出証番号　出証特２００４－３０１６４５８

【書類名】　特許願

【整理番号】　2033840243

【提出日】　平成15年 1月31日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 17/00

G06N 5/00

G06N 7/00

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　森川　幸治

【発明者】

【住所又は居所】　北海道札幌市北区北６条西９丁目　中央公務員宿舎４－１０２

【氏名】　大森　隆司

【発明者】

【住所又は居所】　北海道札幌市北区北１８条西４丁目　シティハイムＮ１８－７０８

【氏名】　大東　優

【発明者】

【住所又は居所】　大阪府門真市大字門真１００６番地　松下電器産業株式会社内

【氏名】　岡　夏樹

【特許出願人】

【識別番号】　000005821

【氏名又は名称】　松下電器産業株式会社

【代理人】

【識別番号】　100077931

【弁理士】

【氏名又は名称】　前田　弘

【選任した代理人】

【識別番号】　100094134

【弁理士】

【氏名又は名称】　小山　廣毅

【選任した代理人】

【識別番号】　100110939

【弁理士】

【氏名又は名称】　竹内　宏

【選任した代理人】

【識別番号】　100110940

【弁理士】

【氏名又は名称】　嶋田　高久

【選任した代理人】

【識別番号】　100113262

【弁理士】

【氏名又は名称】　竹内　祐二

【選任した代理人】

【識別番号】　100115059

【弁理士】

【氏名又は名称】　今江　克実

【選任した代理人】

【識別番号】　100115691

【弁理士】

【氏名又は名称】　藤田　篤史

【選任した代理人】
　【識別番号】　100117581
　【弁理士】
　【氏名又は名称】　二宮　克也
【選任した代理人】
　【識別番号】　100117710
　【弁理士】
　【氏名又は名称】　原田　智雄
【選任した代理人】
　【識別番号】　100121500
　【弁理士】
　【氏名又は名称】　後藤　高志
【選任した代理人】
　【識別番号】　100121728
　【弁理士】
　【氏名又は名称】　井関　勝守
【手数料の表示】
　【予納台帳番号】　014409
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　0217869
【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　予測型行動決定装置および行動決定方法

【特許請求の範囲】

【請求項１】　所定の環境について状態を観察し、状態値を取得する状態観察部と、

前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、

前記環境予測部による予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定する目標状態決定部と、

前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備えた

ことを特徴とする予測型行動決定装置。

【請求項２】　請求項１において、

前記環境予測部は、前記環境の将来の，自己の行動に影響されない状態変化を予測するものである

ことを特徴とする予測型行動決定装置。

【請求項３】　請求項１において、

前記環境の各状態に係る状態価値を記憶する状態価値記憶部を備え、

前記環境予測部は、複数ステップにわたって、将来の状態を予測するものであり、

前記目標状態決定部は、前記環境予測部によって予測された各将来状態について、それぞれ、前記状態価値記憶部を参照して状態価値を求め、目標状態を決定するものである

ことを特徴とする予測型行動決定装置。

【請求項４】　請求項３において、

前記目標状態決定部は、状態価値が極大となる将来状態を、目標状態として決定する

ことを特徴とする予測型行動決定装置。

【請求項５】　請求項３において、

前記目標状態決定部は、前記状態価値記憶部から得た状態価値を、現在からのステップ数に応じて割り引いて、用いる
ことを特徴とする予測型行動決定装置。

【請求項６】　請求項３において、

前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、

当該予測型行動決定装置は、

前記環境予測部によって予測された，自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えたものである
ことを特徴とする予測型行動決定装置。

【請求項７】　請求項１において、

所定の行動基準に基づいて、自己の行動を決定する第２の行動決定部と、

前記第１および第２の行動決定部によって決定された行動を第１および第２の行動候補として受け、これら第１および第２の行動候補のうちのいずれか一方を、実際の行動として選択する行動選択部とを備えた
ことを特徴とする予測型行動決定装置。

【請求項８】　請求項７において、

前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、

前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第１の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第２の行動候補を選択するものである
ことを特徴とする予測型行動決定装置。

【請求項９】　請求項１において、

前記第１の行動決定部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部と、

前記行動付状態変化検出部によって検出された，現在状態並びにその前ステッ

プにおける状態および行動の組合せを、状態変化として記憶する行動付状態変化記憶部と、

前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えたものである

ことを特徴とする予測型行動決定装置。

【請求項１０】　請求項９において、

前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行うものである

ことを特徴とする予測型行動決定装置。

【請求項１１】　請求項１において、

前記環境予測部は、

前記状態値を受け、この状態値が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部と、

前記状態変化検出部によって検出された，現在状態およびその前ステップにおける状態の組合せを、状態変化として記憶する状態変化記憶部と、

前記状態変化記憶部から、現在状態の後の状態を予測する状態予測部とを備えたものである

ことを特徴とする予測型行動決定装置。

【請求項１２】　予測型行動決定装置において、自己の行動を決定する方法であって、

所定の環境について状態を観察して、状態値を取得し、

取得した状態値に基づいて、前記環境の将来の状態変化を予測し、

予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定し、

決定した目標状態を基にして、自己の行動を決定する

ことを特徴とする行動決定方法。

【請求項１３】　請求項１２において、

予測する状態変化は、前記環境の将来の，自己の行動に影響されない状態変化

である

ことを特徴とする行動決定方法。

【請求項１４】　請求項１２において、

目標状態の決定を、予測した各将来状態に係る状態価値を参照して、行う

ことを特徴とする行動決定方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、外部から入力を受け、現在の状態から将来どのような状態に遷移するかを予測しながら、外部への出力を決定する行動決定に関する技術に属する。

【０００２】

【従来の技術】

近年、産業上使用されるシステムは、年々複雑化しており、入力と出力との関係を予めプログラム等によって記述しておくことが困難になりつつある。このため、入力信号を処理して正しい出力を求める方法が必要になっており、このような入力から出力を決定する装置のことを、本願明細書中で「行動決定装置」と呼ぶ。また、特に、入力信号から将来の状態遷移を予測した上で、出力を求めるものを「予測型行動決定装置」と呼ぶ。

【０００３】

行動決定のための従来技術は、１）現在の状態のみによって行動決定を行うもの、２）過去の状態の遷移によって行動決定を行うもの、３）将来の状態を予測して行動決定を行うもの、に分類される。

【０００４】

現在の状態のみによって行動決定を行う従来技術として、ＩＦ－ＴＨＥＮルールを用いるもの、ニューラル・ネットワークを用いるもの、テーブル参照方式などがある。これらは、現在の状態に対する行動が予め記述されており、入力から現在の状態を判別し、所定の記述を参照して行動を決定する。

【０００５】

ところが、現在の状態だけで正しい行動が決定できるとは必ずしも限らない。

例えば対話型ロボットシステムにおいて、「いいですか」と聞かれたとき、それだけでは意味不明であり、それまでの状態遷移を参照することによってはじめて言葉の意味を理解することができる。すなわち、行動決定のために、過去の状態が必要となる場合もある。

【０００６】

また、現在や過去の状態だけでなく、将来の状態も考慮すべき場合もある。例えば、移動ロボットが障害物を回避するような場合には、まだ衝突していない段階では問題は発生していなくて、将来このままの移動方向や速度では衝突するという将来の状態遷移を考慮することによって、はじめて衝突前に回避の行動をとることができる。

【０００７】

将来の状態を考慮した行動決定に関する従来技術として、特許文献１，２に開示された技術がある。特許文献１では、環境から視覚センサや関節角度センサによって得た入力信号から、現在の画像データや関節角度データを状態として受ける。システムは、対象物に対する行動とその行動結果としての画像の変化をリカレント・ニューラル・ネットワークによって記憶し、同じような状況を受けた場合に、記憶している行動を再現する。この技術は、例えばロボットの自律的行動決定に応用されている。また特許文献２では、強化学習における行動決定技術が示されており、ある状態における価値と１ステップ前の状態における価値から、誤差を予測してその情報を行動決定に用いている。

【０００８】

【特許文献１】

特開２００２－５９３８４号公報

【０００９】

【特許文献２】

特開２００２－１８９５０２号公報

【００１０】

【発明が解決しようとする課題】

上述のように、特許文献１では、現在状態から自己の行動によってどのような

状態に変化するかをリカレント・ニューラル・ネットワークを用いて予測している。そして、その予測結果に応じて、状態と対で記憶された行動が決定される。

【００１１】

しかしながら、特許文献１では、自己の行動に起因する過去の状態変化をリカレント・ニューラル・ネットワークを用いて学習しているに過ぎず、自己の行動と関係のない環境の変化については、予測も考慮も何らなされていない。また、ある時点での行動決定は、現在の状態とその１ステップ後の状態予測に基づいて行われているが、この１ステップ後の状態が行動決定のために重要であるとは必ずしもいえず、したがって、この将来の状態予測が行動決定にとって適切なものとはいえない。

【００１２】

また、特許文献２に関しても、現在の状態と１ステップ後の状態の予測価値のみから決定された行動が、必ずしも望ましいものとは限らない。例えば、移動ロボットが走ってくる車をよけたいときに、移動ロボットの速度が車に対してかなり遅い場合は、何ステップも前から回避行動を起こさないと車とぶつかってしまうように、将来を見越して行動を決定すべき場合には、１ステップ後だけでなくさらに将来の状態も考慮して行動を決定する必要がある。また、例えば上述の移動ロボットが走ってくる車をよける場合のように、現在や１ステップ後の価値を見ても変化はみられないが、何ステップも先に決定的な自体が発生するときは、現在や１ステップ後の価値に基づく行動は、無駄な行動につながる可能性もある。

【００１３】

前記の問題に鑑み、本発明は、予測型行動決定装置において、行動決定のために将来の状態予測をより適切に行い、行動決定の精度や能力を向上させることを課題とする。

【００１４】

【課題を解決するための手段】

本発明は、次の点に着目している。すなわち、環境の状態の中には、自己（予測型行動決定装置）の行動（＝出力）に関連するものと、自己の行動に関係なく

変化するものとがあり、前者については、長い将来にわたっての状態予測は困難であるが、その一方で、後者については、１ステップ先だけではなく長い将来にわたっての予測も比較的容易に行うことができる。

【００１５】

すなわち、本発明では、環境を観察して、環境の将来の状態変化を予測し、この予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定する。そして、この決定した目標状態を基にして、自己の行動を決定する。これにより、環境の将来の状態予測により得られた、行動決定のために最も適した目標状態を基にして、行動が決定されるので、従来よりも行動決定の精度が向上する。

【００１６】

【発明の実施の形態】

本発明の第１態様によれば、所定の環境について状態を観察し、状態値を取得する状態観察部と、前記状態観察部によって取得された状態値に基づいて、前記環境の将来の状態変化を予測する環境予測部と、前記環境予測部による予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定する目標状態決定部と、前記目標状態決定部によって決定された目標状態を基にして、自己の行動を決定する第１の行動決定部とを備えた予測型行動決定装置を提供する。

【００１７】

本発明の第２態様によれば、前記環境予測部は、前記環境の将来の自己の行動に影響されない状態変化を予測する第１態様の予測型行動決定装置を提供する。

【００１８】

本発明の第３態様によれば、前記環境の各状態に係る状態価値を記憶する状態価値記憶部を備え、前記環境予測部は、複数ステップにわたって将来の状態を予測するものであり、前記目標状態決定部は、前記環境予測部によって予測された各将来状態について、それぞれ、前記状態価値記憶部を参照して状態価値を求め、目標状態を決定する第１態様の予測型行動決定装置を提供する。

【００１９】

本発明の第４態様によれば、前記目標状態決定部は、状態価値が極大となる将来状態を目標状態として決定する第３態様の予測型行動決定装置を提供する。

【００２０】

本発明の第５態様によれば、前記目標状態決定部は、前記状態価値記憶部から得た状態価値を現在からのステップ数に応じて割り引いて用いる第３態様の予測型行動決定装置を提供する。

【００２１】

本発明の第６態様によれば、前記状態価値記憶部は、自己を含む状態に係る状態価値を記憶するものであり、当該予測型行動決定装置は、前記環境予測部によって予測された，自己を含まない将来状態について、前記状態価値記憶部に記憶された状態価値を基にしてその状態価値を求め、前記目標状態決定部に与える価値変換部を備えた第３態様の予測型行動決定装置を提供する。

【００２２】

本発明の第７態様によれば、所定の行動基準に基づいて自己の行動を決定する第２の行動決定部と、前記第１および第２の行動決定部によって決定された行動を第１および第２の行動候補として受け、これら第１および第２の行動候補のうちのいずれか一方を実際の行動として選択する行動選択部とを備えた第１態様の予測型行動決定装置を提供する。

【００２３】

本発明の第８態様によれば、前記目標状態決定部は、目標状態を決定できたか否かを示す選択信号を前記行動選択部に与えるものであり、前記行動選択部は、前記選択信号が、目標状態を決定できたことを示すときは、前記第１の行動候補を選択する一方、目標状態を決定できなかったことを示すときは、前記第２の行動候補を選択する第７態様の予測型行動決定装置を提供する。

【００２４】

本発明の第９態様によれば、前記第１の行動決定部は、前記状態値を受け、この状態値が表す現在状態からその前ステップにおける状態と行動を検出する行動付状態変化検出部と、前記行動付状態変化検出部によって検出された，現在状態並びにその前ステップにおける状態および行動の組合せを、状態変化として記憶

する行動付状態変化記憶部と、前記行動付状態変化記憶部から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして、行動を決定する行動計画部とを備えた第１態様の予測型行動決定装置を提供する。

【００２５】

本発明の第１０態様によれば、前記行動計画部は、前記状態変化記憶部の検索の際に、目標状態から現在の状態に向かって後ろ向き探索を行う第９態様の予測型行動決定装置を提供する。

【００２６】

本発明の第１１態様によれば、前記環境予測部は、前記状態値を受け、この状態値が表す現在状態からその前ステップにおける状態を検出する状態変化検出部と、前記状態変化検出部によって検出された現在状態およびその前ステップにおける状態の組合せを状態変化として記憶する状態変化記憶部と、前記状態変化記憶部から現在状態の後の状態を予測する状態予測部とを備えた第１態様の予測型行動決定装置を提供する。

【００２７】

本発明の第１２態様によれば、予測型行動決定装置において、自己の行動を決定する方法として、所定の環境について状態を観察して、状態値を取得し、取得した状態値に基づいて、前記環境の将来の状態変化を予測し、予測結果を基にして、行動決定のために最も適した将来の状態を目標状態として決定し、決定した目標状態を基にして、自己の行動を決定する行動決定方法を提供する。

【００２８】

本発明の第１３態様によれば、予測する状態変化は、前記環境の将来の自己の行動に影響されない状態変化である第１２態様の行動決定方法を提供する。

【００２９】

本発明の第１４態様によれば、目標状態の決定を、予測した各将来状態に係る状態価値を参照して行う第１２態様の行動決定方法を提供する。

【００３０】

まず、本発明に関する基本的な概念について説明する。

【００３１】

図１は課題の例を示す図である。図１では、座標（０，０）－（１，１）の空間１において、ボールＢが直進運動をしている。空間１の上下左右の壁に当たったとき、ボールＢは反射する。パドルＰは左右方向にのみ動くことができるものとする。ボールＢの状態は、位置座標（Ｂｘ，Ｂｙ）と進む方向Ｂｔによって表現され、パドルＰの状態は、位置座標（Ｐｘ，Ｐｙ）によって表現される。ただし、Ｐｙ＝０で固定されている。

【００３２】

各時刻ステップにおいて、パドルＰの操作ＭＰとして、｛ＬＥＦＴ（左に動く）、ＲＩＧＨＴ（右に動く）、ＳＴＡＹ（そのまま）｝のうちいずれか１つを選択する。そして、ボールＢをパドルＰによって受けることができたとき、正の報酬が得られるものとし、一方、ボールＢをパドルＰによって受けることができなかったとき、負の報酬が得られるものとする。そしてここでのタスクは、得られる報酬をより多くすることである。

【００３３】

これを行動決定問題として見た場合、各ステップごとに、入力としてボールＢおよびパドルＰの状態（Ｂｘ，Ｂｙ，Ｂｔ，Ｐｘ）が与えられたとき、より多くの報酬が得られるように、パドルＰの操作ＭＰを選択する、ということになる。

【００３４】

このような課題に対して、例えば、状態価値および行動基準を用いて、行動を決定する方法がすでに知られている。このような方法を、本願明細書では「政策（Policy）に基づく行動決定」と呼ぶ。

【００３５】

図２は政策に基づく行動決定に用いられる状態価値と行動基準の例を示す図である。ここでの「状態価値」は、外部の状態を評価して得た価値のことである。図２の例では、図１の空間１が８×８のセルに分割されており、各セルには、ボールＢがその位置にあるときの状態価値が示されている。例えばボールＢがセルＣＬの位置にあるとき、状態価値は「３」である。図２の例では、ボールＢが空間１内の下面に到達したときに報酬がもらえるか否かが決定されることから、下面に近い位置ほど状態価値が高くなっている。状態価値は、事前に与えられたり

、学習によって獲得されたりする。

【００３６】

また、各セルには、その位置に来たときどの行動を取るべきか、という行動基準が記述されている。図２の例では、図１と対応させて考えると、パドルＰがボールＢの下あたりにあり、ボールＢが左下に向かって落ちてきているときに、ボールＢの位置に対応したセルＣＬにおける行動基準として、左に動く行動に対して０．５、動かない行動に対して０．４、右に動く行動に対して０．１、という値が割り当てられている。この値を基にして、例えば最大値をもつ行動を選択したり、行動選択確率として計算したりして、最終的な行動が決定される。例えば最大値をとる行動を選択する場合には、セルＣＬでは、最大値０．５を持つ「左に動く」が選択される。

【００３７】

このように、政策に基づく行動決定では、現在のボールＢの状態（位置）に対応する状態価値が参照され、その状態に対する行動基準によって行動が決定される。すなわち、行動が現在の状態を考慮して決定されるので、例えばボールＢの速度が早い場合など、事前に行動を決定する必要があるような場合等には対応できなかった。

【００３８】

これに対して、本発明では、「予測に基づく行動決定」を行う。すなわち図３（ａ）に示すように、現在のボール位置Ｂ１に対して、複数のステップにわたる将来のボール位置を予測し、最も行動決定に適した目標状態となるボール位置Ｂ２を決定し、この目標状態を基にして行動決定を行う。

【００３９】

図３（ｂ）は予測された各ボール位置に対する状態価値が格納されたテーブル、図３（ｃ）は予測ステップ数に対する状態価値の変化を示すグラフである。図３（ｃ）から分かるように、予測ステップ数を増やすにつれて、ボールＢは下面に向かって移動することになり、これとともに状態価値は徐々に増加する。そして、下面に到達してからはボールＢは上方に向かうため、状態価値は減少に転じる。このため、状態価値は、予測ステップ数５のとき、最大値「８」になる。こ

こで、例えば最大の状態価値を取った予測ステップ数と予測状態を行動決定に用いるものとすると、位置Ｂ２が目標状態となる。これは、ボールＢの状態でいうと、一番下面に近いときが目標状態として設定されることに相当する。このように、将来の，自己の行動に影響されない状態変化を予測して、予測した状態の中で行動決定のために最も適したものを目標状態として設定して行動決定を行うことによって、現在の行動をより的確に決定することができる。

【００４０】

以下、本発明の実施の形態について、図面を参照して説明する。なお、以下の説明では、図１の課題をタスクとして扱うものとする。図１の課題は、ピンポンのようにボールが移動してそれを跳ね返すと報酬が得られるものであるが、これは、サッカーゲームにおいてボールを取りに行く行動、移動ロボットが接近物を回避する行動等と同様の設定とみなすことができる。接近物の回避では、接近物の動きを予測してこれを回避することを目標状態として設定することによって、行動を決定できる。この他にも、ロボットが物を受け取る際の行動決定なども、同様の課題となる。

【００４１】

（第１の実施形態）

図４は本発明の第１の実施形態に係る予測型行動決定装置１０の構成を示す。この予測型行動決定装置１０は、所定の環境１１について状態を観察し、環境１１の，自己の行動に影響されない状態変化を予測した上で、目標状態を定め、環境１１に対する行動を決定する。

【００４２】

状態観察部１２は環境１１の状態を観察し、現在の状態を表す状態値を取得する。ここでは、状態値をｓ（ｔ）と表す。図１の課題については、ボールＢおよびパドルＰの状態（Ｂｘ，Ｂｙ，Ｂｔ，Ｐｘ）が状態値ｓ（ｔ）として得られる。

【００４３】

環境予測部１３は状態観察部１２によって取得された状態値ｓ（ｔ）に基づいて、環境１１の将来の状態変化を予測する。ここでは、ボールＢの座標（Ｂｘ，

Ｂｙ）を状態ｓ’として予測するものとする。すなわち、ボールＢは基本的には直進し、壁で反射するので、将来の座標は、現在のボールＢの座標と角度から解析的な計算によって予測することができる。またこの予測には、パドルＰの操作ＭＰを考慮する必要はない。すなわち、当該装置１０の行動に影響されない、将来の状態変化を予測することになる。本実施形態では、環境予測部１３はこのような予測機能を予め備えているものとする。なお、このような予測機能は学習等によって得ることも可能であり、これについては後述する。

【００４４】

状態価値記憶部１４は図３（ｂ）に示すように、各状態ｓ’すなわちボールの位置（Ｂｘ，Ｂｙ）について状態価値をそれぞれ格納している。ここでは、状態ｓ’に対する状態価値をＶ（ｓ’）と記述する。本課題では、ボールＢが下の壁に当たるとき、報酬がもらえるか否かが判定されるので、下の壁に近い位置ほど、より高い状態価値が設定されている。なお、状態価値は全ての状態について設定されているのが好ましいが、これは実際には困難なので、例えば特定の部分状態にのみ状態価値を設定しておいてもよい。

【００４５】

また、状態ｓ’の要素にボールＢの進む方向Ｂｔを含めてもよく、この場合は、状態価値記憶部１４に、（Ｂｘ，Ｂｙ，Ｂｔ）について状態価値をそれぞれ格納しておけばよい。

【００４６】

目標状態決定部１５は、環境予測部１３による予測結果を基にして、行動の決定のために最も適した将来の状態を目標状態として決定する。ここでは、環境予測部１３によって複数ステップにわたって予測された将来の状態ｓ’（ｔ＋Δ）について、それぞれ、状態価値記憶部１４を参照して状態価値Ｖ（ｓ’（ｔ＋Δ））を求め、求めた状態価値から目標状態を決定する。

【００４７】

第１の行動決定部としての予測に基づく行動決定部１６は、目標状態決定部１５によって決定された目標状態に対して、現在どのような行動を取るべきかを決定する。ここでは、ボールＢの座標とパドルＰの座標が一致したとき報酬がもら

えるので、目標状態から、パドルＰがどの位置に向かえば報酬をもらえる可能性が高くなるを識別し、その位置に向かうように、操作ＭＰを決定する。

【００４８】

以下、図５のフローチャートを参照して、図４の予測型行動決定装置１０の動作を説明する。まず、状態観察部１２によって、環境１１から、現在の状態を表す状態値ｓ（ｔ）が取得される（Ｓ１１）。そして、何ステップ先の予測を行うかを変数Δで定義し、その初期値として１を与える（Ｓ１２，Ｓ１３）。なお、ステップＳ１７で所定の条件を満たすまで、または、Δが所定値ｎを超えるまで（Ｓ１９）、Δをインクリメントしながら、以下のステップＳ１４～Ｓ１６を繰り返し実行する。

【００４９】

ステップＳ１４において、環境予測部１３は、現在状態からΔステップ先の状態ｓ’（ｔ＋Δ）を予測する。例えば、ボールＢが壁に当たらないときは、Ｂｘ，Ｂｙが１ステップ分変化し、方向Ｂｔは変化しない。そしてステップＳ１５において、状態価値記憶部１４から、ステップＳ１４で予測された状態ｓ’（ｔ＋Δ）の状態価値Ｖ（ｓ’（ｔ＋Δ））が抽出される。

【００５０】

ステップＳ１６において、目標状態決定部１５は、状態価値記憶部１４から出力された状態価値Ｖ（ｓ’（ｔ＋Δ））を評価する。ここでは、評価条件として、所定の閾値を超えているか否かを判定する。そして、条件を満たす、すなわち状態価値が所定の閾値を越えていると判定したときは（Ｓ１７でＹＥＳ）、ステップＳ１８に進み、状態ｓ’（ｔ＋Δ）を目標状態として、予測に基づく行動決定部１６に与える。

【００５１】

一方、条件を満たさないときは（Ｓ１７でＮＯ）ステップＳ１９に進み、予測ステップ数の判定を行う。すなわち、Δが所定値ｎを下回るときは（Ｓ１９でＮＯ）ステップＳ１３に戻り、Δをインクリメントした後、同様の処理を行う。一方、Δが所定値以上のときは（Ｓ１９でＹＥＳ）ステップＳ１Ａに進み、目標状態が決定できなかった旨を予測に基づく行動決定部１６に通知する。

【００５２】

予測に基づく行動決定部１６は、目標状態決定部１５からの出力を受けて、次の行動を決定する。しかし、目標状態が決定できなかった通知を受けた場合は、例えばランダムに、行動を決定する。

【００５３】

このように本実施形態によると、環境１１の状態変化の予測結果から、状態価値を参照して、行動決定のために最も適した将来の状態が目標状態として決定される。そして、この目標状態を基にして、自己の行動が決定される。このため、従来の将来を予測した行動決定よりも、行動決定の精度が格段に向上する。また、従来技術のように状態と行動との関係を予め記述していなくても、行動が決定できるので、簡易な構成により、様々な未経験の事態に対しても行動決定可能になる。また、環境１１の将来の，当該装置１０の行動に影響されない状態変化を予測しているので、１ステップ先だけではなく長い将来にわたっての予測も比較的容易に、精度良く、行うことができる。

【００５４】

なお、本実施形態では、状態価値が所定の閾値を越えたときの状態を目標状態として決定するものとしたが、目標状態の決定は、これ以外にも、様々な方法が考えられる。例えば、予測したステップ内で状態価値が最大となった状態や、状態価値が極大になった状態を、目標状態として決定してもよい。また、前ステップとの状態価値の差分が所定値よりも小さくなったときの状態を目標状態として決定してもよい。あるいは、現在からのステップ数に応じて、状態価値を割り引いて評価するようにしてもよい。

【００５５】

なお、本実施形態では、状態価値がテーブル形式で記憶されているものとしたが、この代わりに、ニューラル・ネットワーク等の関数近似手法を適用してもよい。この場合、現在の状態が入力されたとき、報酬の期待値が出力されるように学習がなされている必要がある。

【００５６】

（第２の実施形態）

図６は本発明の第２の実施形態に係る予測型行動決定装置１０Ａの構成を示すブロック図である。図６において、図４と共通の構成要素には図４と同一の符号を付している。

【００５７】

まず、価値変換部２１について説明する。価値変換部２１は、状態価値記憶部１４Ａが当該装置１０Ａの行動によって変化する状態を含む状態に係る状態価値を記憶しているとき、環境予測部１３によって予測された，当該装置１０Ａの行動によって変化する状態を含まない将来状態ｓ’（ｔ＋Δ）の状態価値Ｖ（ｓ’（ｔ＋Δ））を、状態価値記憶部１４Ａに格納された状態価値を基にして求め、目標状態決定部１５Ａに与える。本実施形態では、状態価値記憶部１４Ａは、パドルＰの位置Ｐｘも含めた状態（Ｂｘ，Ｂｙ，Ｂｔ，Ｐｘ）についてそれぞれ状態価値を記憶しているものとし、環境予測部１３からは将来状態（Ｂｘ，Ｂｙ，Ｂｔ）が出力されるものとする。

【００５８】

図７のフローチャートを参照して、価値変換部２１の動作を説明する。まず、環境予測部１３によって将来状態ｓ’（ｔ＋Δ）が予測されたとき、状態価値記憶部１４Ａをこの状態ｓ’（ｔ＋Δ）によって検索する（Ｓ２１）。そして、３個の要素（Ｂｘ，Ｂｙ，Ｂｔ）が一致する状態について、その状態価値の組Ｖ’（ｓ’（ｔ＋Δ））を抽出する（Ｓ２２）。ここで、パドルＰの座標Ｐｘが８通りであるとすると、８個の状態価値が探索結果として出力される。

【００５９】

そして、ステップＳ２２において抽出された状態価値を比較し、その最大値を求め（Ｓ２３）、得られた最大値を状態ｓ’（ｔ＋Δ）の状態価値Ｖ（ｓ’（ｔ＋Δ）として目標状態決定部１５Ａに出力する。

【００６０】

また、第１の実施形態では、状態価値記憶部１４に格納された状態価値は事前に与えられるものとしたが、本実施形態では、強化学習と呼ばれる手法によって自動的に学習させるものとする。「強化学習」とは、参考文献１（サットン（R. S. Sutton），バート（A.Barto）著，「強化学習入門（Reinforcement Learnin

g: An Introduction）」，（米国），A Bradford Book, The MIT Press, １９９８年３月）で知られる学習方法の１種であり、環境からの報酬信号と試行錯誤によって学習を行う手法であり、報酬の最大化を目的として行動決定を行う学習アルゴリズムのことである。

【００６１】

本実施形態では、強化学習の中で、ａｃｔｏｒ－ｃｒｉｔｉｃ学習という手法を利用する（参考文献１のｐ．１５１～１５３参照）。ａｃｔｏｒは、どの状態でどの行動を取るべきかを記述する行動決定方策ｐｏｌｉｃｙを持ち、状態ｓで、とりうる行動ａ１～ａｎに対して選択確率が算出されて決定される。ｃｒｉｔｉｃのＶには状態価値と呼ばれる、ある状態でどれぐらい報酬がもらえそうかの期待値を示す値が格納される。状態価値は、報酬から計算されるＴＤ誤差によって更新される。ＴＤ誤差δは、

$\delta = r(t+1) + \gamma(V(s(t+1)) - V(s(t))$

で計算される。ここでｒ（ｔ＋１）は報酬、γは割引率（discount rate）と呼ばれている。このＴＤ誤差によって、ｃｒｉｔｉｃ内部の状態価値テーブルＶ（ｓ）と、ａｃｔｏｒの持つＰｏｌｉｃｙの更新を行う。

【００６２】

ａｃｔｏｒ－ｃｒｉｔｉｃ法を利用して、状態価値記憶部１４Ａに記憶される状態価値を更新することによって、状態価値が不明な状態であっても、予測によって行動を決定することができる。状態価値更新部２２はｃｒｉｔｉｃテーブル更新則によって、状態価値記憶部１４Ａ内に記憶された状態価値を更新する。

【００６３】

図８は強化学習によって学習された状態価値記憶部１４Ａの状態価値に基づき、価値変換部２１が求めた状態価値の変化を示すグラフである。図８において、縦軸は算出された状態価値、横軸は予測のステップ数である。図８のグラフでは、予測ステップ数の増加に伴って状態価値が増加しており、８０ステップあたりでピークを迎えている。一般に、状態価値は、図８のように、未学習領域から既学習領域に入ると増加し始め、その後に減少に転じる。この減少に転じるタイミングが、現在状態から予測されるもっとも意義のある状態、すなわち目標状態に

相当すると判断できる。

【００６４】

目標状態設定部１５Ａは、第１の実施形態と同様に動作する。ただし、目標状態を決定できたか否かを示す選択信号を、行動選択部２４に与える点が異なっている。予測に基づく行動決定部１６Ａは、第１の実施形態と同様に動作して行動を決定し、決定した行動を第１の行動候補として行動選択部２４に出力する。

【００６５】

第２の行動決定部としての政策に基づく行動決定部２３は、所定の行動基準としての、強化学習におけるａｃｔｏｒ－ｃｒｉｔｉｃ手法によって学習された政策に基づいて行動を決定し、決定した行動を第２の行動候補として行動選択部２４に出力する。

【００６６】

行動選択部２４は、予測に基づく行動決定部１６Ａから受けた第１の行動候補と、政策に基づく行動決定部２３から受けた第２の行動候補とのうち、いずれか一方を、環境１１への実際の行動として選択する。この選択には、目標状態決定部１５Ａから受けた選択信号を利用する。すなわち、選択信号が目標状態を決定できなかったことを示すときは、予測に基づく行動決定部１６Ａから受けた第１の行動候補は意味がないと考えて、政策に基づく行動決定部２３から受けた第２の行動候補を実際の行動として選択する。そうでない場合は、予測に基づく行動決定部１６Ａから受けた第１の行動候補を実際の行動として選択する。

【００６７】

このように本実施形態によると、状態価値を強化学習によって求めることによって、あらかじめ与えることが困難であった状態価値が自律的に獲得され、予測型行動決定が容易に実現できる。また、学習装置としての観点から本装置を見た場合は、報酬が与えられる時点を予測することになり、初期の実学習の状態が多い場合にも行動の指針を得ることができ、学習の効率が向上する。

【００６８】

以下、従来の強化学習（Reinforcement Learning；以下、「ＲＬ」と記載）と、本実施形態に係る学習方法（Prediction-based Reinforcement Learning; 以

下「ＰＲＬ」と記載)のシミュレーションによる比較結果をいくつか示す。このシミュレーションでは、ボールが打てたときの報酬は１．０、打てなかったときの報酬は－１．０、またパドルが右または左に動いたときの報酬は－０．０１とした。また、ＰＲＬでは、最初に状態価値記憶部１４Ａ内の状態価値をある程度作成するために３０００試行（ボールが下面に当たるまでを１試行といい、以下epochと記載）は、ＲＬと同様の処理で行動決定を行った。各epochにおけるボールの初期位置はランダムに設定した。

【００６９】

図１３はＲＬとＰＲＬそれぞれの累積報酬の変化を表すグラフである。横軸はepoch数、縦軸は累積報酬である。最初の３０００epochでは、両手法の累積報酬にはほとんど差はない。これは、この期間では、ＰＲＬはＲＬと同じ行動決定方法と学習法で動いているからである。学習の３０００epoch以降では、ＰＲＬの方が、環境予測部１３と予測に基づく行動決定部１６Ａとの利用によって、ＲＬよりも良い性能を示している。これは、ＰＲＬでは、まだ学習が収束していない（すなわち、状態価値記憶部１４Ａ内の状態価値が完全には生成されていない）ときでも、予測によって適切な行動が決定できるので、ボールを打ち返す確率がＲＬよりも高くなるからである。また、ＰＲＬはボールを打ち返す位置の予測に基づき行動を決定しているため、不必要な行動をする割合が低くなっていることもその原因の一つと考えられる。

【００７０】

図１４は学習１００epoch毎に評価を行った結果を示すグラフである。横軸はepoch数、縦軸は１００epochごとに行われた性能評価の結果であり、ボールが打てた割合である。ここでのＲＬの行動決定規則には、ソフトマックス（softmax）と呼ばれる行動戦略を用いている。これは、各状態の各行動に付与された値に応じて確率的に行動決定を行う方法である。初期の３０００epochの学習後、ＰＲＬmodelのパフォーマンスが、ＲＬと比べて大きく向上している。この結果は、まだボールを打ち返す付近の状態しか学習が進んでいない（すなわち、状態価値が決定されていない）ＲＬ学習の初期段階に、ＰＲＬが環境予測部１３と予測に基づく行動決定部１６Ａとの利用によって、学習が進行していない状態に対して

も適切な行動を決定できたことが大きな要因である。

【００７１】

図１５はＲＬの行動決定規則をgreedy戦略に変更した場合のグラフである。図１４と比べて、ＲＬの性能は全体的に見ると向上している。しかし、学習していない状態に評価時に遭遇すると性能は急激に落ちる。ＲＬは、ＰＲＬに比べると、精度が悪く安定した動作を示していないことが分かる。

【００７２】

なお、本実施形態では、状態価値をテーブルとして表現して強化学習により学習するものとしたが、ニューラル・ネットワークによって学習させることもできる。この場合、ニューラル・ネットワークの汎化能力によって未経験の状態に対する価値を出力することが期待される。しかし、この手法は、状態価値に不連続な状態が少ないときに有効であると考えられる。

【００７３】

（第３の実施形態）

上述の実施形態では、予測に基づく行動決定部１６，１６Ａの機能は予め与えられているものとしたが、行動生成機能を予め与えることが難しい場合は、目標状態に到達するための行動生成能力の獲得が必要になる。本実施形態では、このような行動生成能力を学習によって獲得させるものとする。

【００７４】

図９は図６の構成における予測に基づく行動決定部１６Ａの本実施形態に係る内部構成を示す図である。図９において、３１は状態値ｓ（ｔ）を受け、この状態値ｓ（ｔ）が表す現在状態から、その前ステップにおける状態と行動を検出する行動付状態変化検出部、３２は行動付状態変化検出部３１によって検出された現在状態ｓ（ｔ）並びにその前ステップにおける状態ｓ（ｔ－１）および行動ａ（ｔ－１）の組合せを状態変化として記憶する行動付状態変化記憶部、３３は行動付状態変化記憶部３２から、現在状態から目標状態までの期間の状態変化の履歴を検索し、この検索結果を基にして行動を決定する行動計画部である。

【００７５】

図１０のフローチャートを参照して、図９の予測に基づく行動決定部１６Ａの

動作を説明する。図１０（ａ）は状態変化を蓄積する場合の動作、図１０（ｂ）は行動計画を行う場合の動作を示す。この２つの動作は、同時並行して実行できる。

【００７６】

状態変化を蓄積する場合、まず、環境１１から状態ｓ（ｔ）を受ける（Ｓ３１）。この現在の状態ｓ（ｔ）とそのときの行動ａ（ｔ）はワーキングメモリに蓄えておく。そして、前ステップにおける状態ｓ（ｔ－１）と行動ａ（ｔ－１）をワーキングメモリから取り出し（Ｓ３２）、状態ｓ（ｔ）とともに行動付状態変化記憶部３２に格納する（Ｓ３３）。これは、状態ｓ（ｔ－１）のときに行動ａ（ｔ－１）をとったら状態ｓ（ｔ）に変化したという、行動に起因する状態変化を表している。

【００７７】

また、行動計画を行う場合、まず、目標状態決定部１５Ａから送られてきた目標状態を探索したい状態ｘｓ（ｎ）として設定する（Ｓ３４）。そして、探索したい状態ｘｓ（ｎ）を行動付状態変化記憶部３２から探索し（Ｓ３５）、検出されたときは（Ｓ３６でＹＥＳ）、探索したい状態ｘｓ（ｎ）と対になって記憶されている状態ｘｓ（ｎ－１）および行動ｘａ（ｎ－１）をワーキングメモリに格納する（Ｓ３７）。その後、ステップＳ３８に進み、探索したい状態の１ステップ前の状態ｘｓ（ｎ－１）が現在の状態ｓ（ｔ－１）でなければ（Ｓ３８でＮＯ）、探索したい状態を更新し（Ｓ３９）、ステップＳ３５に戻る。同様の処理を繰り返し実行し、ステップＳ３８において、探索したい状態の１ステップ前の状態ｘｓ（ｎ－１）が現在の状態ｓ（ｔ－１）と一致したとき（ＹＥＳ）、それまでにワーキングメモリに格納されていた状態ｘｓと行動ｘａの系列を行動計画として出力する。

【００７８】

一方、ステップＳ３６において、探索したい状態ｘｓ（ｎ）が行動付状態変化記憶部３２から検出できないときは（ＮＯ）、行動不能と判断して（Ｓ３Ａ）、処理を終了する。なおこの場合は、予測に基づく行動決定部１６Ａから、正しい行動が決定できないという信号が出力され、行動選択部２４は、政策に基づく行

動決定部２３から出力された第２の行動候補を実際の行動として選択する。

【００７９】

このような動作によって、現在の行動のみならず、現在状態から目標状態に至るまでの行動計画が得られるので、行動計画が一旦完了した後は、その行動計画に従って行動候補を順に出力すればよい。これにより、処理量が格段に少なくなるので、特に長期にわたる予測誤差が少ない場合は、好ましい。もちろん、毎ステップごとに、目標状態までの行動計画を算出し直してもよい。この場合は、予測が完全でない場合であっても、行動が決定できる。

【００８０】

なお、本実施形態では、目標状態から現在状態に向かって後ろ向き探索を行うものとしたが、現在の状態ｓ（ｔ）と行動ａ（ｔ）からｓ（ｔ＋１）を算出する前向き探索を用いても、同様に行動計画を作成することができる。

【００８１】

（第４の実施形態）

本実施形態では、学習によって状態予測を実現するものとする。

【００８２】

図１１は図４および図６の構成における環境予測部１３の本実施形態に係る内部構成を示す図である。図１１において、４１は状態値ｓ（ｔ）を受け、この状態値ｓ（ｔ）が表す現在状態から、その前ステップにおける状態を検出する状態変化検出部、４２は状態変化検出部４１によって検出された現在状態ｓ（ｔ）およびその前ステップにおける状態ｓ（ｔ－１）を状態変化として記憶する状態変化記憶部、４３は状態変化記憶部４２から、現在状態の後の状態を予測する状態予測部である。

【００８３】

図１２のフローチャートを参照して、図１１の環境予測部１３の動作を説明する。図１２（ａ）は状態変化を蓄積する場合の動作、図１２（ｂ）は状態予測を行う場合の動作を示す。この２つの動作は、同時並行して実行できる。

【００８４】

状態変化を蓄積する場合、まず、環境１０から状態ｓ（ｔ）を受ける（Ｓ４１

）。この現在の状態ｓ（ｔ）はワーキングメモリに蓄えておく。そして、前ステップにおける状態ｓ（ｔ－１）の組合せをワーキングメモリから取り出し（Ｓ４２）、状態ｓ（ｔ）とともに状態変化記憶部４２に格納する。これは、状態ｓ（ｔ－１）の次に状態ｓ（ｔ）になったという状態変化を表している。

【００８５】

また、状態予測を行う場合、まず、環境１１から取得した現在の状態ｓ（ｔ）を探索したい状態ｙｓ（ｎ）として設定する（Ｓ４４）。そして、探索したい状態ｙｓ（ｎ）を状態変化記憶部４２から探索し（Ｓ４５）、検出されたときは（Ｓ４６でＹＥＳ）、探索したい状態ｙｓ（ｎ）と対になって記憶されている１ステップ後の状態ｙｓ（ｎ＋１）を状態変化記憶部４２から取り出し、出力する（Ｓ４７）。その後、ステップＳ４８に進み、目標状態決定部１５，１５Ａから評価信号による再予測依頼を受けたときは（ＹＥＳ）、探索したい状態を更新し（Ｓ４９）、ステップＳ４５に戻る。

【００８６】

一方、ステップＳ４６において、探索したい状態ｙｓ（ｎ）が状態変化記憶部４２から検出できないときは（ＮＯ）、予測先不明と判断して（Ｓ４Ａ）、処理を終了する。なおこの場合は、環境予測部１３から、正しい予測ができないという信号が出力され、行動選択部２４は、政策に基づく行動決定部２３から出力された第２の行動候補を実際の行動として選択する。

【００８７】

このような方法によって、環境予測部１３の機能を予め作成しなくても、学習によって取得することができる。

【００８８】

なお、状態変化記憶部４２の学習のために、ニューラル・ネットワーク等の関数近似手法によって次の状態を予測させることもできる。この場合、ニューラル・ネットワークが本来持つ汎化能力によって、経験したことがない状態ｓ（ｔ）に対しても、適切な１ステップ語の状態ｓ（ｔ＋１）を出力できる可能性がある。

【００８９】

【発明の効果】

以上のように本発明によると、環境の将来状態を予測し、予測の結果決定された、行動決定のために最も適した目標状態を基にして、行動が決定されるので、将来の状態変化がより適切に考慮され、行動決定の精度が向上する。

【図面の簡単な説明】

【図１】

本発明、および各実施形態を説明するための課題を示す図である。

【図２】

政策に基づく行動決定に用いられる状態価値と行動基準の例である。

【図３】

本発明に係る，予測に基づく行動決定を示す図である。

【図４】

本発明の第１の実施形態に係る予測型行動決定装置の構成を示すブロック図である。

【図５】

図４の予測型行動決定装置の動作を示すフローチャートである。

【図６】

本発明の第２の実施形態に係る予測型行動決定装置の構成を示すブロック図である。

【図７】

図６の構成における価値変換部の動作を示すフローチャートである。

【図８】

予測ステップ数に伴う状態価値の変化の一例を示すグラフである。

【図９】

本発明の第３の実施形態における予測に基づく行動決定部の内部構成を示す図である。

【図１０】

図９の構成の動作を示すフローチャートである。

【図１１】

本発明の第４の実施形態における環境予測部の内部構成を示す図である。

【図１２】

図１１の構成の動作を示すフローチャートである。

【図１３】

本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

【図１４】

本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

【図１５】

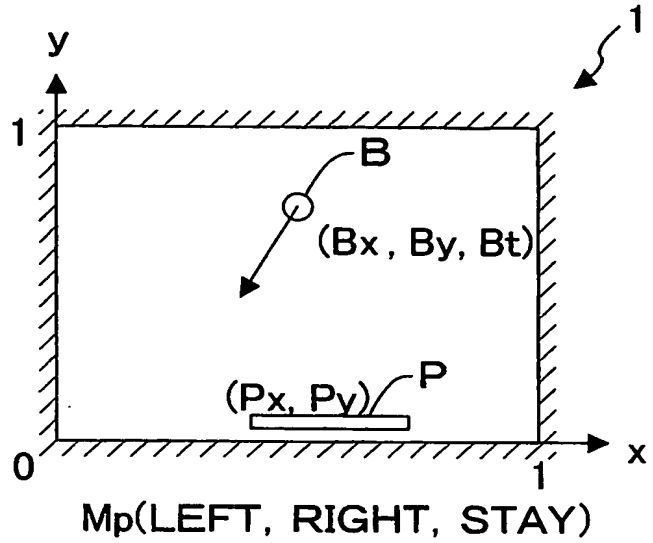
本発明の実施形態に係る学習方法と従来の強化学習のシミュレーション結果を示すグラフである。

【符号の説明】

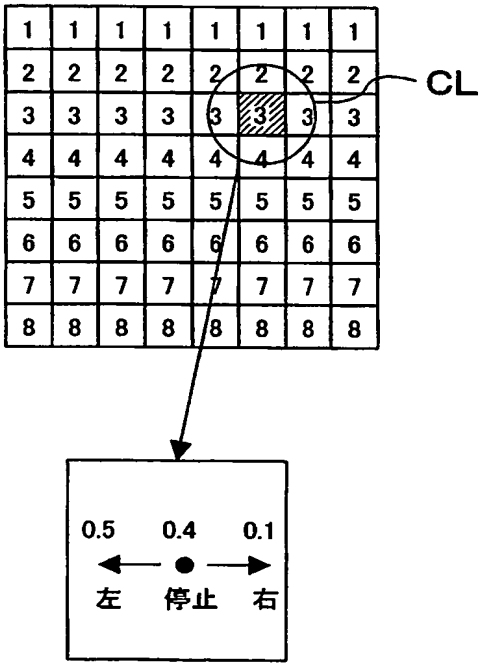
１０，１０Ａ　予測型行動決定装置
１１　環境
１２　状態観察部
１３　環境予測部
１４，１４Ａ　状態価値記憶部
１５，１５Ａ　目標状態決定部
１６，１６Ａ　予測に基づく行動決定部（第１の行動決定部）
２１　価値変換部
２３　政策に基づく行動決定部（第２の行動決定部）
２４　行動選択部
３１　行動付状態変化検出部
３２　行動付状態変化記憶部
３３　行動計画部
４１　状態変化検出部
４２　状態変化記憶部
４３　状態予測部

【書類名】　図面

【図１】

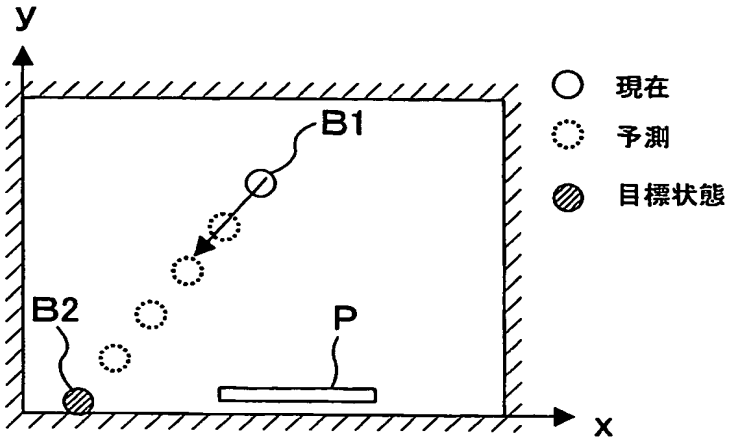


【図２】

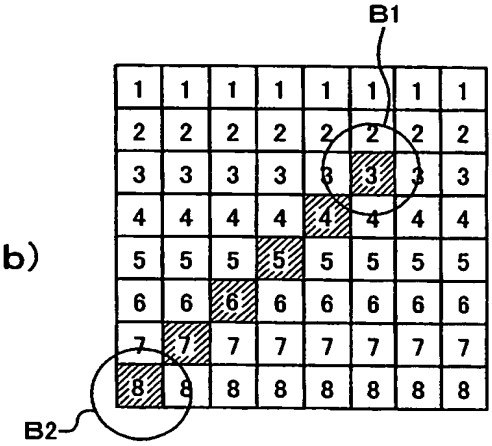


【図３】

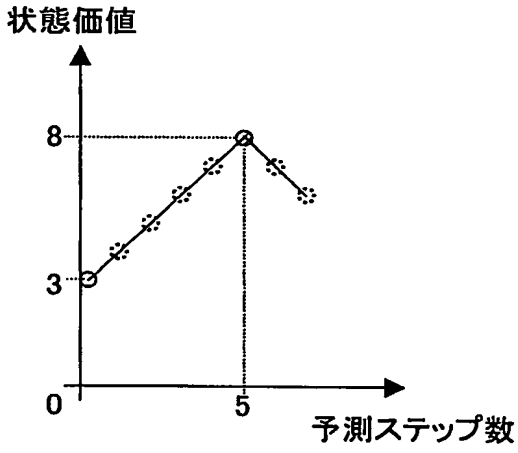
(a)



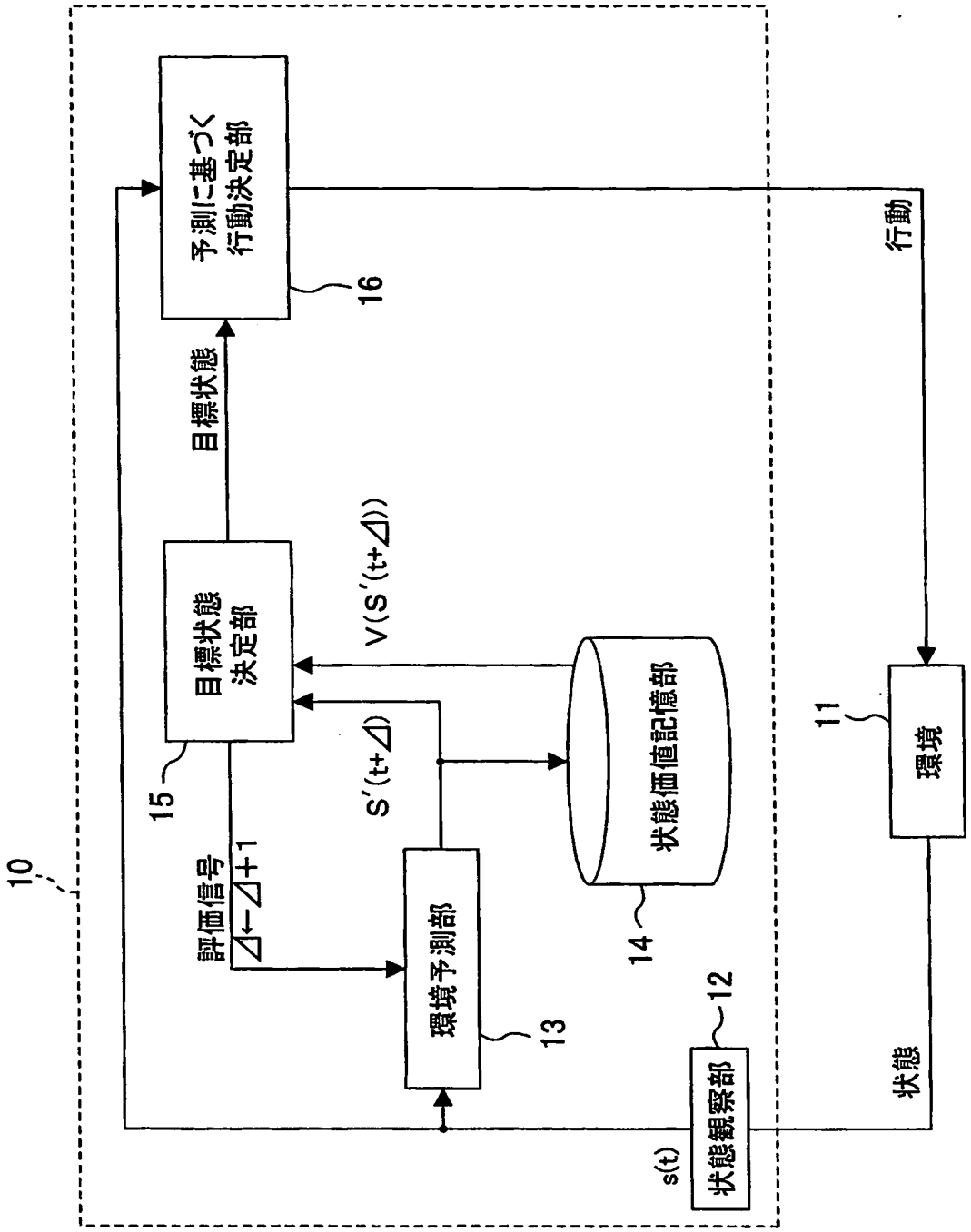
(b)



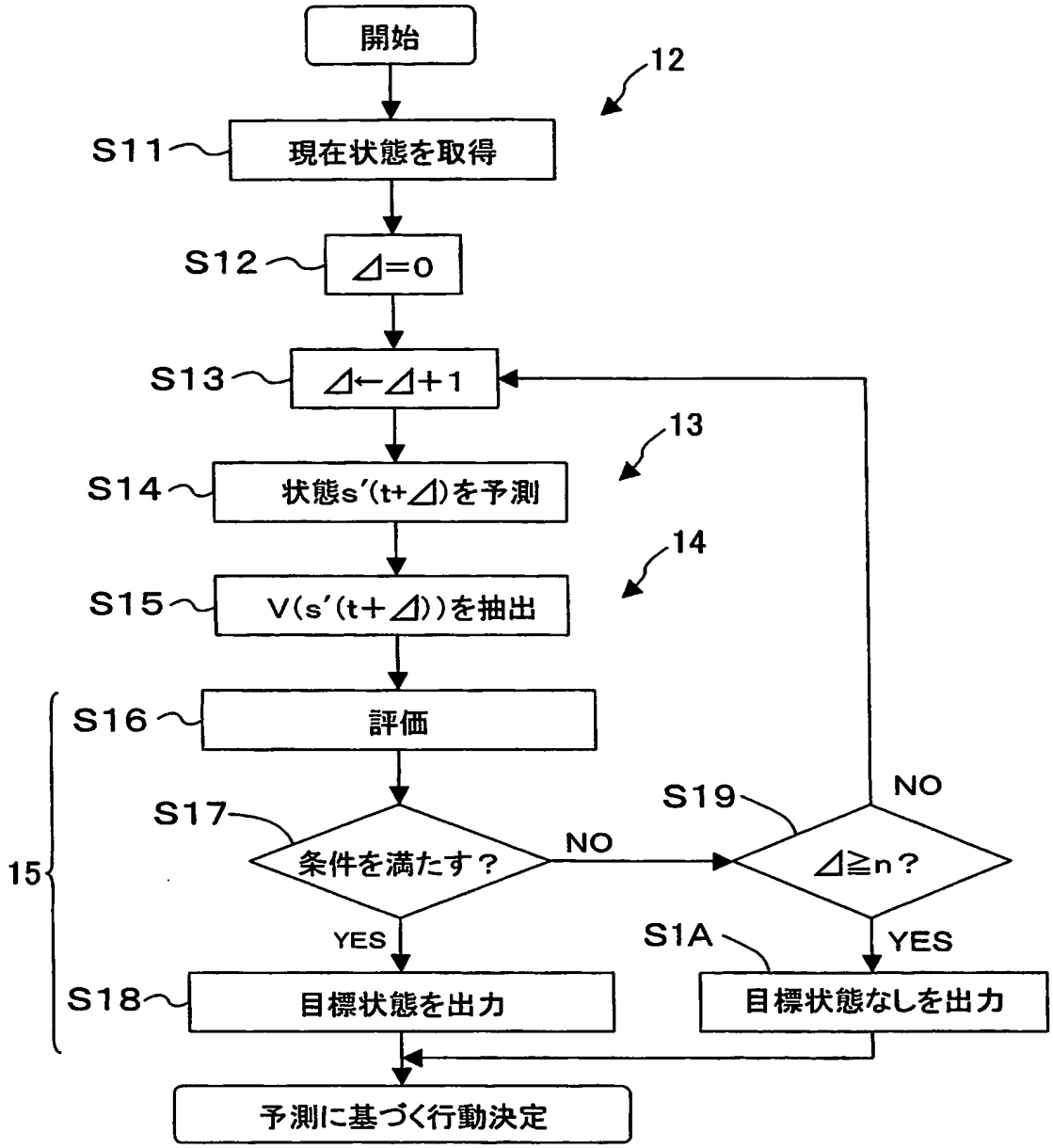
(c)



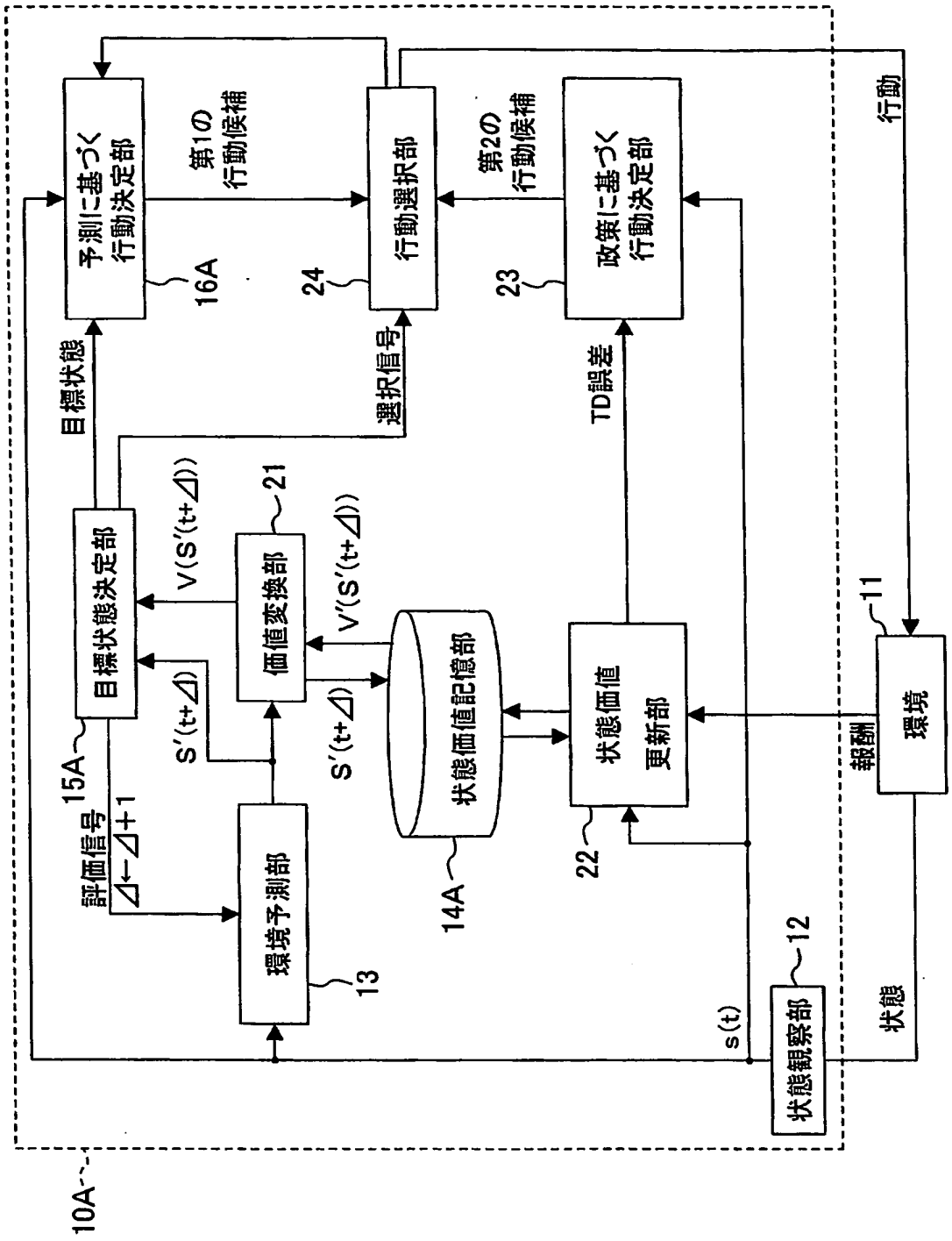
【図４】


予測に基づく行動決定部
16
行動
目標状態
目標状態決定部
15
V(S'(t+Δ))
S'(t+Δ)
評価信号
Δ←Δ+1
環境予測部
13
状態価値記憶部
14
11
環境
10
12
状態観察部
状態
s(t)

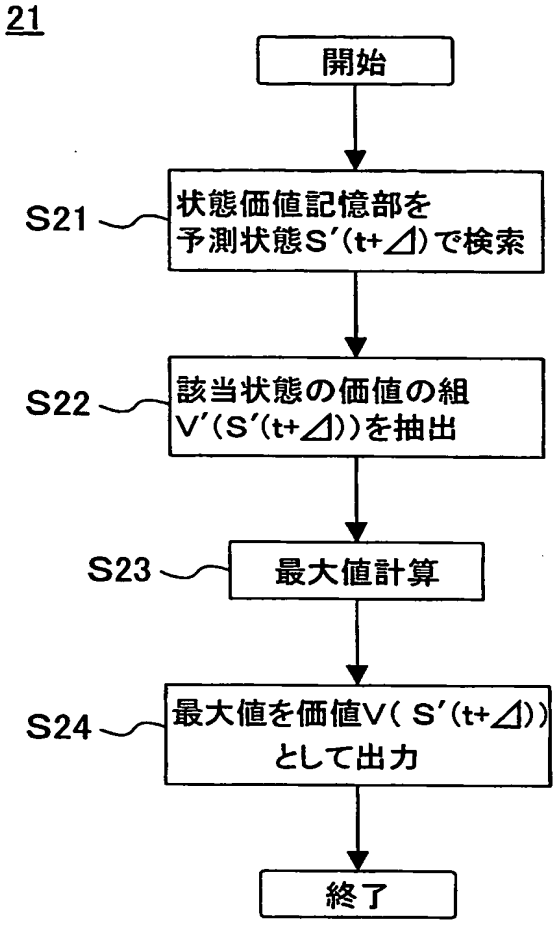
【図５】


開始
12
S11
現在状態を取得
S12
⊿=0
S13
⊿←⊿+1
13
S14
状態s'(t+⊿)を予測
14
S15
V(s'(t+⊿))を抽出
S16
評価
S17
条件を満たす？
NO
S19
⊿≧n？
NO
15
YES
S1A
YES
S18
目標状態を出力
目標状態なしを出力
予測に基づく行動決定

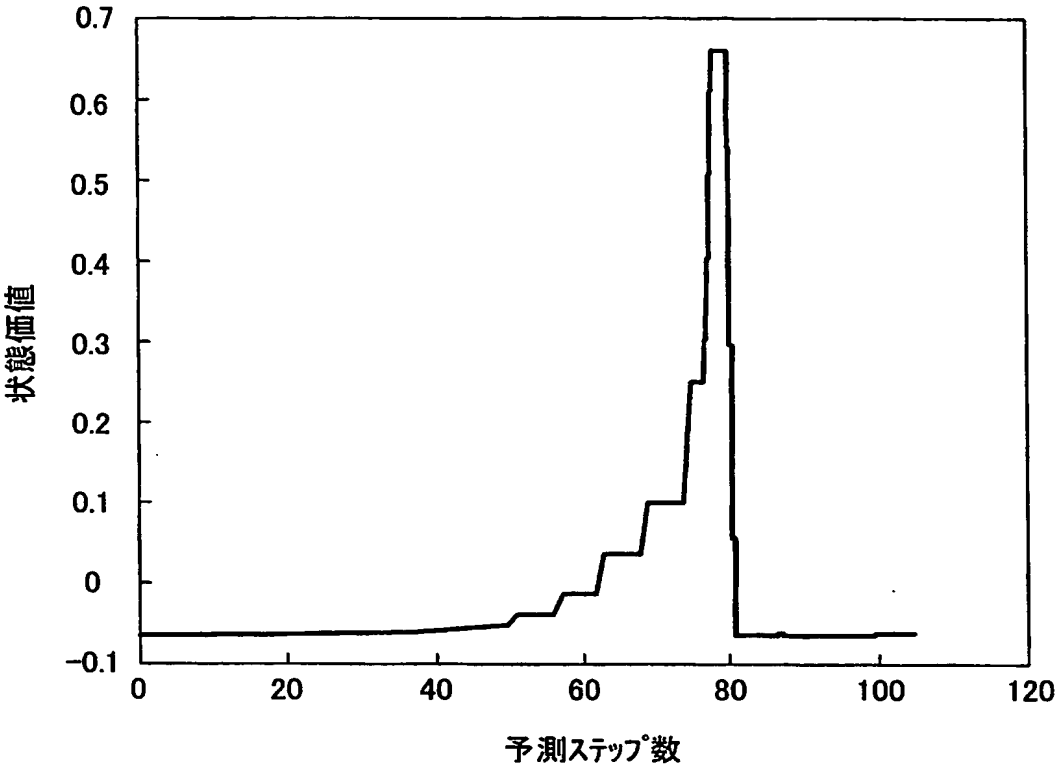
【図６】


予測に基づく行動決定部
第1の行動候補
行動選択部
第2の行動候補
政策に基づく行動決定部
行動
16A
24
23
目標状態
選択信号
TD誤差
目標状態決定部
V(S'(t+Δ))
価値変換部
21
V'(S'(t+Δ))
11
15A
S'(t+Δ)
S'(t+Δ)
状態価値記憶部
状態価値更新部
環境
報酬
評価信号
Δ←Δ+1
環境予測部
13
14A
22
12
s(t)
状態観察部
状態
10A

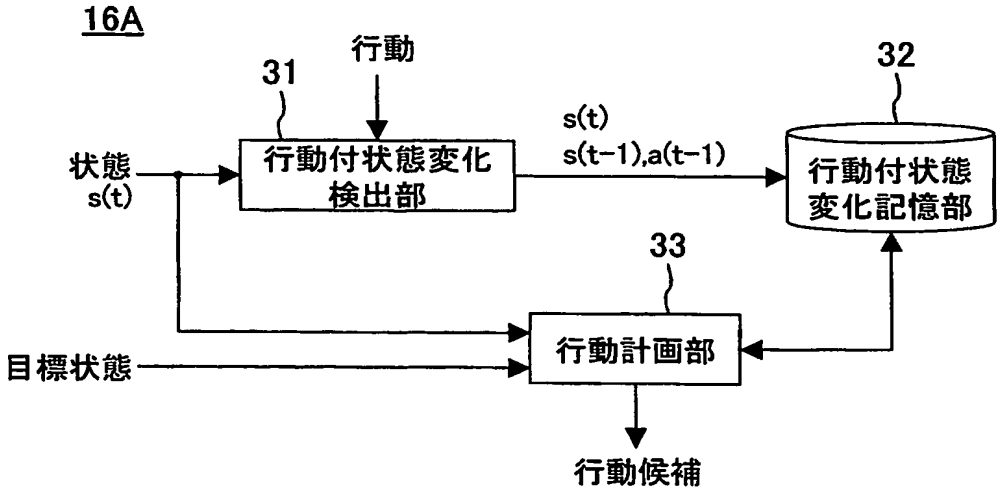
【図７】


21
開始
S21
状態価値記憶部を
予測状態S′(t+⊿)で検索
S22
該当状態の価値の組
V′(S′(t+⊿))を抽出
S23
最大値計算
S24
最大値を価値V(S′(t+⊿))
として出力
終了

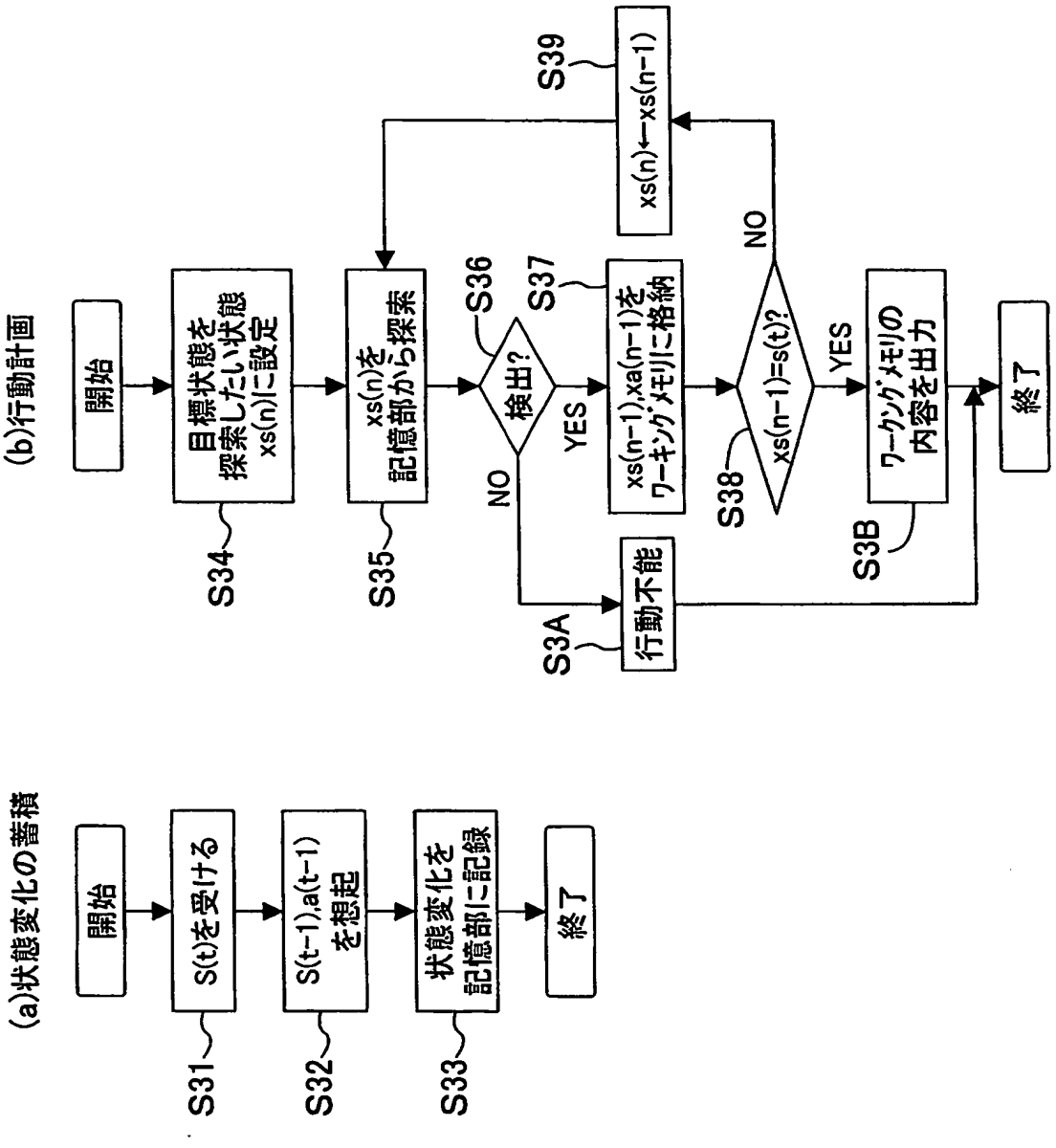
【図８】


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−0.1
状態価値
0
20
40
60
80
100
120
予測ステップ数

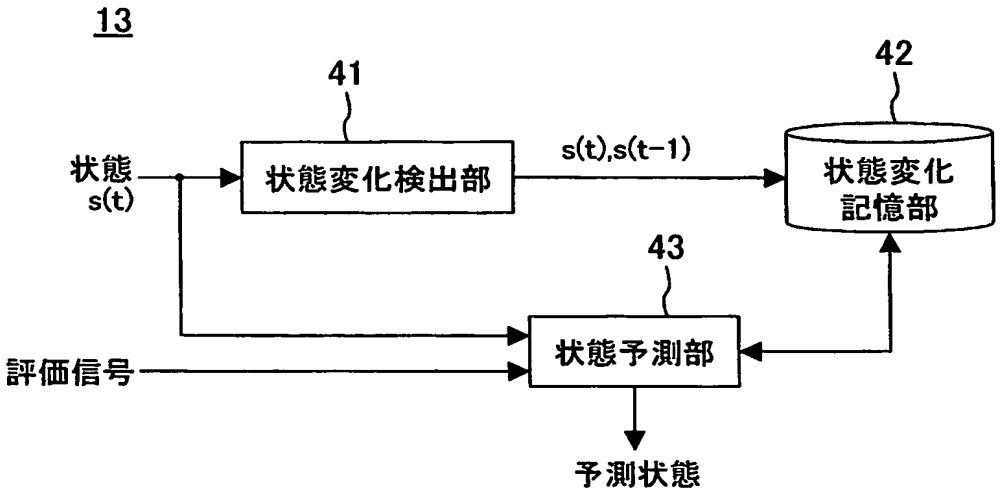
【図９】


16A
行動
31
32
s(t)
s(t−1),a(t−1)
状態
s(t)
行動付状態変化
検出部
行動付状態
変化記憶部
33
行動計画部
目標状態
行動候補

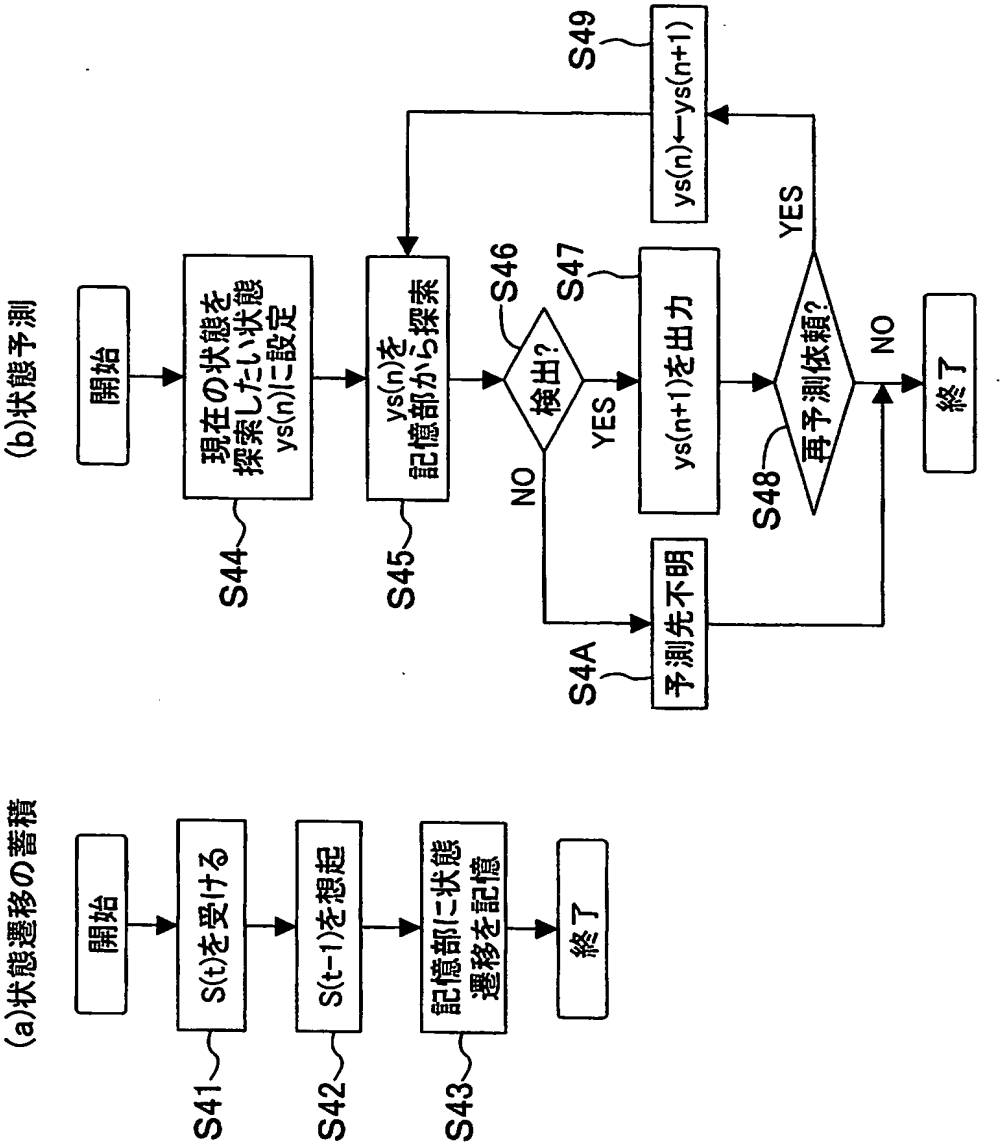
【図１０】


(b)行動計画
開始
目標状態を
探索したい状態
xs(n)に設定
S34
xs(n)を
記憶部から探索
S35
S36
検出?
NO
YES
S37
xs(n-1),xa(n-1)を
ワーキングメモリに格納
S38
xs(n-1)=s(t)?
NO
YES
S39
xs(n)←xs(n-1)
S3A
行動不能
S3B
ワーキングメモリの
内容を出力
終了
(a)状態変化の蓄積
開始
S(t)を受ける
S31
S(t-1),a(t-1)
を想起
S32
状態変化を
記憶部に記録
S33
終了

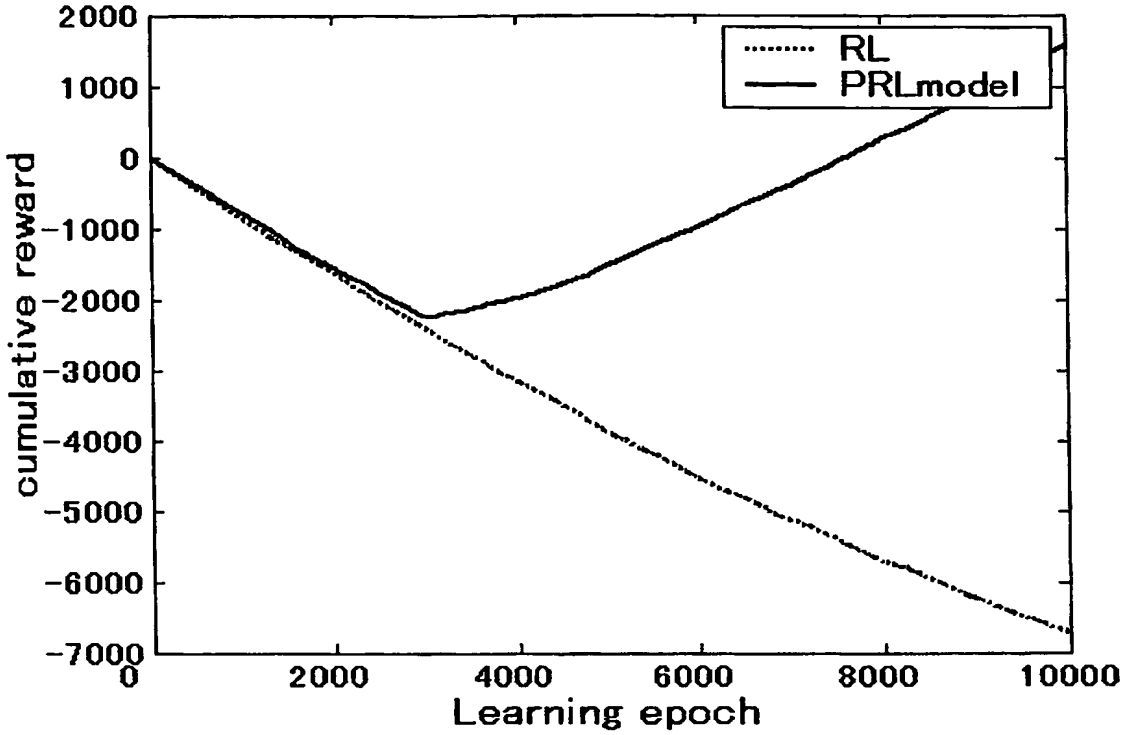
【図１１】



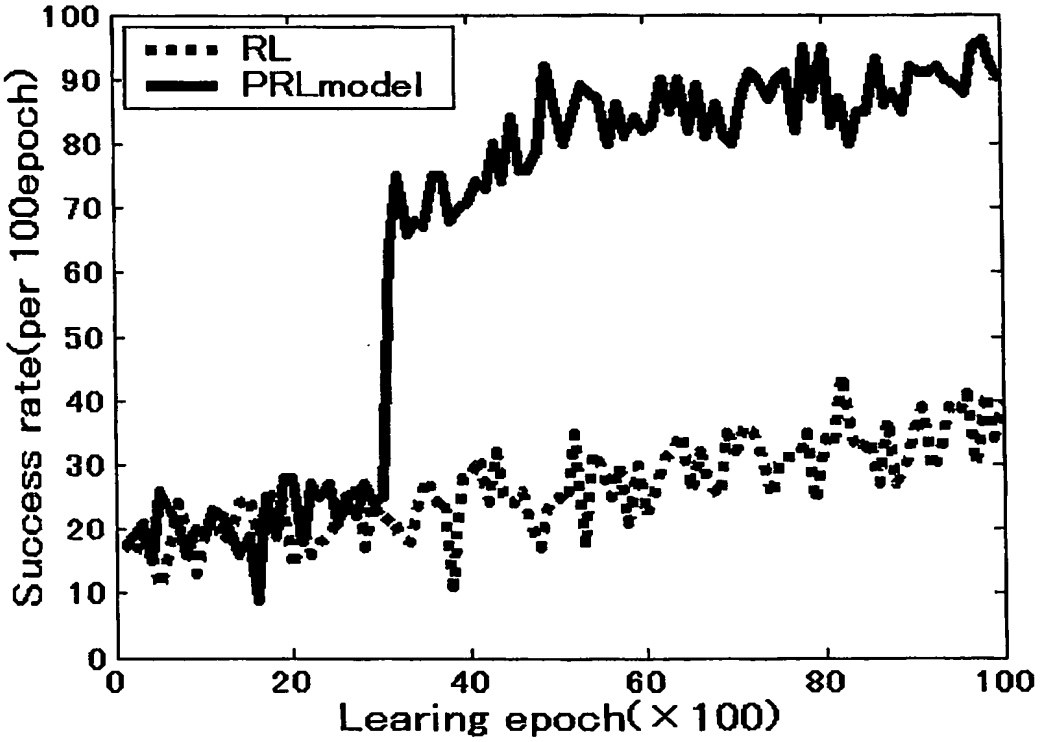
【図１２】


(b)状態予測
開始
現在の状態を探索したい状態ys(n)に設定
S44
ys(n)を記憶部から探索
S45
S46
検出?
NO
YES
S47
ys(n+1)を出力
予測先不明
S4A
S48
再予測依頼?
YES
NO
S49
ys(n)←ys(n+1)
終了
(a)状態遷移の蓄積
開始
S(t)を受ける
S41
S(t-1)を想起
S42
記憶部に状態遷移を記憶
S43
終了

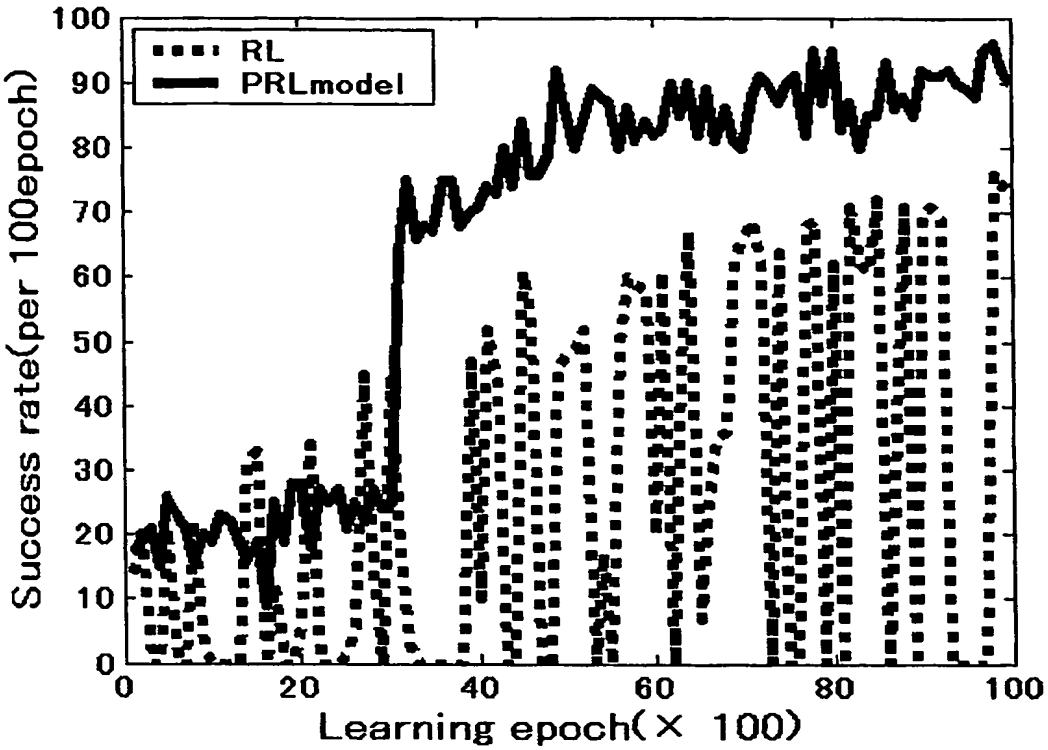
【図１３】


RL
PRLmodel
cumulative reward
Learning epoch
2000
1000
0
-1000
-2000
-3000
-4000
-5000
-6000
-7000
0
2000
4000
6000
8000
10000

【図１４】


RL
PRLmodel
Success rate(per 100epoch)
Learing epoch(×100)
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
100

【図１５】


RL
PRLmodel
Success rate(per 100epoch)
Learning epoch(× 100)
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100

【書類名】　要約書

【要約】

【課題】　予測型行動決定装置において、行動決定のために将来の状態予測をより適切に行い、行動決定の精度や能力を向上させる。

【解決手段】　予測型行動決定装置１０において、状態観察部１２は所定の環境１１について状態を観察し、状態値ｓ（ｔ）を取得する。環境予測部１３は取得された状態値ｓ（ｔ）に基づいて、環境１１の将来の，装置１０の行動に影響されない状態変化を予測する。目標状態決定部１５は状態価値記憶部１４を参照して、行動決定のために最も適した将来の状態を目標状態として決定する。予測に基づく行動決定部１６は決定された目標状態を基にして、自己の行動を決定する。

【選択図】　図４

特願２００３－０２３３４２

出　願　人　履　歴　情　報

識別番号　［０００００５８２１］

１．変更年月日　１９９０年　８月２８日
［変更理由］　新規登録
住　所　大阪府門真市大字門真１００６番地
氏　名　松下電器産業株式会社